Применение обучения с подкреплением в многопользовательских стратегических играх

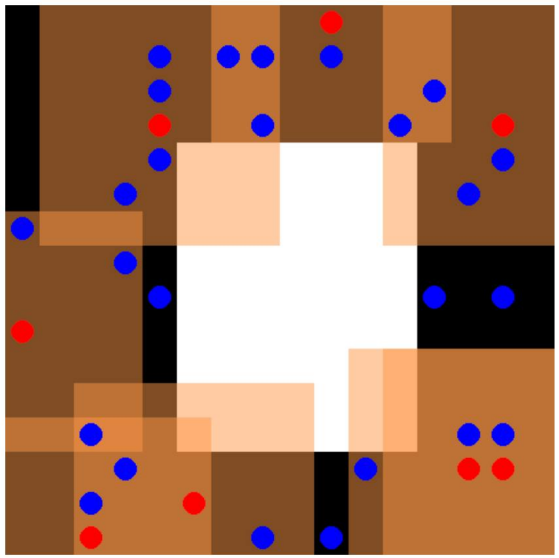
Мисник А.С.
misnik.as@phystech.edu

Центр когнитивного моделирования МФТИ

2025 – Научно-исследовательский семинар

Введение

- Модели многопользовательских игр раскрывают особенности кооперативного взаимодействия
- Выбор оптимальных действий в условиях кооперативного взаимодействия представляет особый интерес для практических применений



Пример модели многопользовательской игры - Pursuit

Мотивация

- Получение структурного представления эталонных политик в обучении с подкреплением
- Обобщение полученного результата - формирование набора человекочитаемых правил
- Следование принципам объяснимого ИИ - обязательный критерий использования ИИ в некоторых сферах

Постановка задачи

S - пространство состояний среды, A - пространство действий
$P(s'|s,a)$ - вероятности переходов состояний при выполнении действия
$R(s,a)$ - награда за выполнение действия a в состоянии s
$\gamma \in [0,1]$ — коэффициент дисконтирования

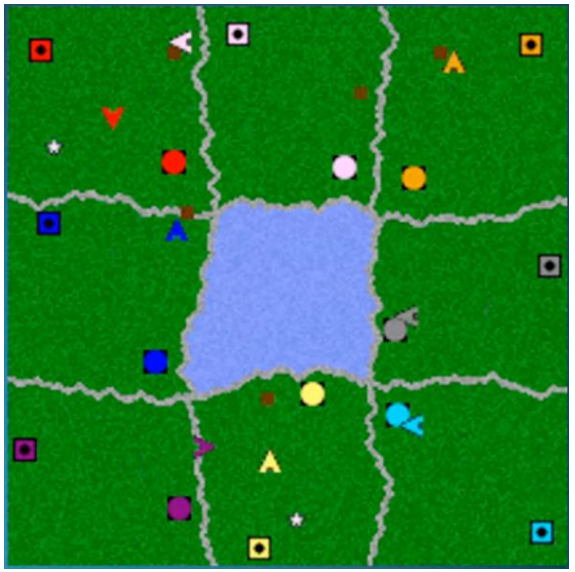
- Цель агента в классическом RL - нахождение оптимальной политики $\pi(a|s)$, определяющей выбор действий в состояниях:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \right]$$

- Выберем набор бинарных признаков исходных состояний и зафиксируем отображение
  F $(\forall s \in S : F(s) = \{f_1, f_2, .., f_n\}, f_i \in \{0,1\})$
  Цель агента в модели задачи - порождение множества гипотез H $(\forall h_i \in H : h_i = \{F_i, A_i\}, F_i \subset \{0,1\}^n)$, где каждая гипотеза имеет свойство $\forall s_t \in S : F(s_t) \in F_h \Rightarrow A_h \sim \pi(\cdot|s_t)$

Модель рассматриваемой игры

1. Игра состоит из тактов, все агенты совершают действия одновременно
2. На каждом из тактов с некоторой вероятностью появляется ресурс, за переработку которого агент получает награду
3. После переработки ресурса появляются отходы, которые возможно утилизировать
4. Количество отходов на своем участке и на участках соседей влияет на вероятность появления новых ресурсов
5. Наблюдение агентов - неполный фрагмент среды, окружающей их



Подзадачи работы

- Разбор нейросетевых подходов и их особенностей
- Создание среды визуализации последовательности фреймов
- Интеграция ВКФ в среду игры
- Эксперименты с порождением гипотез и сравнение результатов

Эксперименты (VDN)

VDN[1] - алгоритм кооперативного обучения с подкреплением, основанный на разложении общей награды в сумму индивидуальных наград.

Проблемы рассмотренного подхода:

- VDN действует в предположении, что все агенты управляются пользователем
- Сформированные стратегии являются сугубо кооперативными

[1] P Sunehag et al., "Value-decomposition networks for cooperative multi-agent learning", 2017

Эксперименты (MAPPO)

MAPPO[2] - адаптация алгоритма Proximal Policy Optimization для многоагентного обучения.

Особенности:

- Требует больше итераций обучения (и хуже параллелится) в сравнении с VDN
- Требует тонкой настройки гиперпараметров

[2] C Yu et al., "The surprising effectiveness of ppo in cooperative multi-agent games", 2022

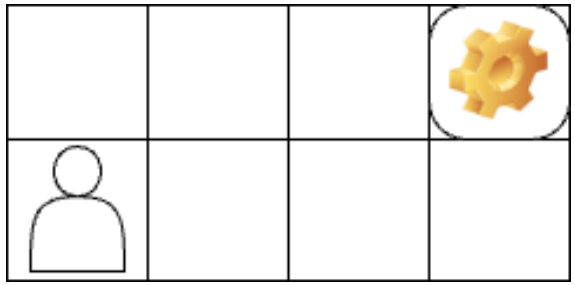
Сравнение подходов

	Score (8 agents)	Score (4 agents vs 4 foreign agents)
VDN	42.53	37.11
MAPPO	46.94	42.03

- Из результатов видно, что оба метода работают хуже при наличии агентов с отличающимися стратегиями
- Даже после визуализации игрового процесса невозможно корректно "объяснить" стратегию

Структурное обучение с использованием ВКФ

Сбор структурной информации о состояниях позволяет агенту мыслить в терминах "сценариев", которые однозначно разбиваются в последовательность действий.



Пример сценария, который разбивается в последовательность (UP R R R PICK)

Проблемы:

- Сложность формирования набор бинарных параметров состояний среды, достаточных для обучения
- Существование состояний, не покрытых гипотезами

Дальнейшие эксперименты будут направлены на обработку этих проблем.

Спасибо за внимание!

misnik.as@phystech.edu